Ridge Regression based Missing Data Estimation with Dimensionality Reduction: Microarray Gene Expression Data

K. Ashfaq Ahmed
Research Scholar, Department of Computer Science & Engineering, Acharya Nagarjuna University, Guntur, A.P, India.
E-mail: kashfaqahmed@gmail.com, ashfaqme@gmail.com

Dr. Shaheda Akthar
Faculty of Computer Science, Government College for Women(A), Guntur, A.P, India.
E-mail: shahedaakthar76@gmail.com

Received September 23, 2021; Accepted December 18, 2021
ISSN: 1735-188X
DOI: 10.14704/WEB/V19I1/WEB19271

Abstract

Data is considered to be the important element in the field of Data Science and Machine Learning. Performance of Machine Learning and Data Mining algorithms greatly influenced by the characteristics of data and data with missing values. Performance of all these Machine Learning algorithms greatly improved and they can give accurate results when the data is in full without missing values. So before applying these algorithms; dataset and its missing values are completely filled. To impute these missing values in the dataset there are numerous methods were proposed. In this paper we used micro array gene expression dataset; by introducing various percentages of missing values a new methodology is proposed to impute these missing values in the data set. The nature of micro array gene expression dataset is huge in dimensionality, so at first, we used recursive feature elimination method to select the best features which contributes much for model was selected then we apply the Ridge Regression for imputation. Imputations with other methods are compared. We evaluate the performance of all models by using the metrics like MSE, MAE, R-square. To select the best model in the set of models we used Normalized Criteria Distance (NCD) to rank the models under proposed metrics. The model with least NCD rank selected as the best model among other models, in our paper proposed model has got the lowest value among other models and considered to be the best model among other models.

Keywords

Missing Data, Micro Array Gene Expression Data, Ridge Regression, Dimensionality Reduction, Normalized Criteria Distance.

Introduction

Development and advancement of Science and Technology in each field has lead to increase amount of data. The evolved vast data is useful for extracting the knowledge, making analysis and visualization for decision. Data mining is a technique used for generating useful information from the data. These datasets are incomplete form with missing values in it. Algorithms perform well only with complete datasets. Depending on the data environment and domain missing of values are introduced in the datasets. Data with missing values can makes the performance of data mining techniques worse and some time with no use.

$$MicroArray\ data = D_{ij} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & . & . & a_{1n} \\ a_{21} & a_{22} & a_{23} & . & . & NA \\ NA & a_{32} & NA & . & . & a_{3n} \\ . & . & . & . & NA & . \\ . & . & . & . & . & . \\ a_{m1} & NA & ? & . & . & a_{mn} \end{pmatrix} \quad (1)$$

Micro array gene expression data is a two-dimensional matrix. Here from the expression (1) D_{ij} represents the gene expression data matrix where subscript 'i' represents the genes and 'j' represents the experimental condition under given environmental condition. Gene expression matrix consisting of values which represents the corresponding gene excitation value under different conditions. From the matrix D_{ij} few of the values are missing represented by NA which represents the missing data values in gene expression data. NA values are distributed randomly in row or column of the dataset randomly.

Motivation behind this Work

Acuna et al., (2004) used Linear Discernment Analysis and KNN Classifier on missing value data and predicted the performance of these algorithms using imputation techniques such as mean, median, and KNN. Missing data values have been shown to have an impact on data mining algorithms such as classification, prediction, and estimation by Brown et al., (2003). On missing values in datasets imputed with mean, Hot Deck, and Navie Byes, Farhangfar et al., (2008) conducted a performance analysis of various data mining classification techniques such as Naive Bayes, RIPPER, C4.5, KNN, and Support Vector Machine. Eekhout et al., (2012) proposed that missing values be replaced by the mean of observed values, which is useful for univariate analysis but not so much for multivariate analysis because this type of imputation method is biased. The discussion on random forest was led by Breiman (2001), who demonstrated that the main advantage of this algorithm is that it can run in parallel, saving computational time. During the data mining process, Y.A.C, (2012) conducted a detailed analysis of the negative impact of missing values in

data sets. All data sets and their correlation among the data are identified and later used to predict missing values by Troyanskaya et al., (2001).

Singular value decomposition and Bayesian Principal Component Analysis are two examples. Depending on the situation and needs, datasets can have one, two, three, or more dimensions. Microarray gene expression data, which is obtained through the gene experimental process, is one type of large dataset. Typically, microarray gene expression data is used to identify a person's medical characteristics. These higher-dimensional data are analysed using algorithms. The raw data of DNA and RNA nucleotides is used to extract gene expression data. The raw DNA nucleotide is mixed with acid and applied to the glass plates as an array of micro spots in two colours: red and green. Before the images are captured, the glass plates are properly dried and washed. Under the microscope, images are captured, and these images are then processed to obtain microarray gene expression data. These captured microarray gene expressions is in the form of matrix array where the columns represents the experimental conditions and rows represent the genes. Process used is hybridization and there are two types in hybridization, one southern blotting, and one northern blotting. In Southern blotting, a small string of positive DNA is hybridized with a complementary segment of DNA and will undergo a process called electrophoresis. In the Northern blotting, Radio labeled DNA hybridized with messenger RNA.

Rubin (1976), Hoheisel et al., (2006), Armstong et al., (2002) Probes with radio material are exposed to different intensities of light when the plate is exposed to fluorescent light. An image is created from the emitted light of various intensities. The obtained image is then subjected to image processing techniques in order to determine the location of dots and their intensities. The obtained intensities are organised into a table, and the values are normalised. Microarrays with complete values are required for gene expression analysis. If there are some missing values or entries in the microarray data, the analysis result may not be satisfactory. These missing entries are frequently found in microarray data due to dust particles and improper image capture. As a result, data with missing entries must be estimated and imputed to improve the accuracy of the results after the analysis. Rubin (1976), Little et al., (2014) These missing entries are classified as MCAR (Missing Completely at Random), MAR (Missing at Random), and MNAR (Missing Completely at Random) (Missing Not at Random). MCAR is a missing data pattern in which the missing value pattern is independent of both observed and unseen values. Missing data is calculated using MCAR, which is completely dependent on external sources. The MAR pattern of missing or missing data is entirely dependent on the observed values. The unobserved distribution of values determines the missing pattern of data in MNAR. Missing data estimate algorithms use three main techniques: local approach, global approach, and a

hybrid strategy that uses both local and global approaches. The local method takes into consideration the dataset's local data pattern of observed values. The global method is based on algorithms that take into consideration the full dataset's global information matrix. Yoon et al., (2007) introduced Robust least square estimation by extending conventional least square estimation with principal component analysis and using quantile regression to estimate missing data. Celton et al., (2010) conducted a comparison investigation of several imputation techniques, evaluating their effectiveness using two datasets: yeast and human gene expressions. Li et al., (2015) presented a hybrid technique called recursive mutual imputation, which combines global dataset correlation with local least square and Bayesian principal component analysis.

The aim of research is to apply the Ridge regression algorithm and with minimizing the dimensionality of data. In the proposed method we use Ridge egression for imputation and recursive feature elimination (RFE) for dimensionality reduction by culling the felicitous features. Section III describes about real time datasets used for analysis. Section IV provides frame work and technical information on dimensionality reduction with RFE. Section V presents related works. Section VI provides the detailed information on proposed model. Section VII presents metrics used to analyze the performance of various models. Section VIII presents results of various models. Section VI is on Conclusion.

Data Sets Used

To analyze the performance of proposed model and previous models we used the following real time Microarray gene expression datasets from previous studies and published in research journal papers. Table I shows the size of the datasets and their corresponding features.

Table I (Different Datasets used for study)

S. No	Dataset Name	No of Rows	No Of Columns
1	Spellman (Spellman et al., (2013))	4381	23
2	Colon Tumor (Alon et al., (1999))	2000	60

From the TABLE I the dataset Spellman consisting of 4381 rows and 23 columns and Colon Tumor dataset consisting of 2000 rows and 60 columns.

Dimensionality Reduction

Larger dimensionality is an issue while we are using some prediction algorithms. Dataset consisting of large number of features can reduce the performance of learning algorithms. It is therefore we first apply to our proposed model. We use the recursive feature elimination

where the least important features are pruned from current set of features the data. Table III shows the datasets used in this paper along with features before and after feature elimination.

Recursive Feature Elimination (Kuhn et al., (2019)

This is a consequential method for culling the consequential features from the datasets. This method propagates as it is facilely configurable. While configuring this method requires how many numbers of consequential features required and cull of the algorithm. Performance of this algorithm entirely depends on these hyper-parameters. This method can be useful in both classification and regression algorithms. This algorithm utilizes the wrapper type technique to extract the features from the pristine dataset. Wrapper means, this algorithm takes the advantage of another machine learning algorithm for feature selection. In this experiment we used AdaBoostRegressor as wrapper algorithm.

Recursive Feature Extraction Algorithm (Hastie et al., 2017)

1. Train the model with all the features with AdaBoost Regressor
2. Now compute the performance of model
3. Calculate the features importance and arrange them in descending order of their importance
4. Eliminate the features with least important
5. Repeat the step 1 to 4 until model reached to best performance and left with important best features.

Table III (Number of features left after RFE)

S no	Dataset Name	No of Columns	No of Columns after RFE
1	Spellman	23	15
2	Colon Tumor	60	20

TABLE III describes the elimination in number of features. This algorithm starts with all the features in the dataset and recursively eliminates the less paramount features while retaining the paramount features.

Related Work

Several studies and methods have been proposed for different types of imputation techniques for missing values. In this paper for a comparative analysis we use existing methods and tools for estimating missing values in the dataset. We discuss briefly some of the methods used in estimating the missing values in data.

Linear SVR [Awad et al., (2015)]

Support vector are one of the greatest techniques which are useful for classification and regression problems. Effectively though the support vector is popular for the classification by formulating convex optimization, it also has great advantage in the prediction with regression. The optimization problem in the classification is to find maximum extent of boundaries that separate the hyperplane. Support vector regression (SVR) is generalizations of classification problem where there is need to find the region of hyperplane with respect to error sensitive region. Support vector regression depends on the optimal hyperplane region for continuous valued function. Support Vector Machines estimates by controlling the model complexity and its prediction error. The shape of hyperplane and its boundaries resembles a cylindrical shape, which is optimized by considering all the training dataset.

$$f(x) = w^T x + b \quad x, w \in R^n \tag{2}$$

The f(x) is a continue values function and the data is a multidimensional data where x and w represents the instances and weights. The function f(x) is linear regression line which is represented in one dimensional hyperplane with equation represents the multivariate regression equation.

Bayesian Ridge [Bishop (2006)]

The difference between the regression and Bayesian regression is the way they take care of coefficients. Usually in general regression coefficients are assigned with single value after the training whereas Bayesian ridge regression, coefficients values as estimated with distribution with certain mean and variance. Bayesian ridge regression generally assigns some prior distribution values to coefficients and reaches to the values of posterior distribution after the training data sets through Bayes theorem.

Decision Tree Regressor [Breiman et al., (1984)]

Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed.

Boot strap aggregation or bagging is considered to be ensemble algorithms which greatly improves the model performance by minimizing the bias and controlling the variance. In the random forest an individual tree is built on the random sample of datasets approximately two third of the total datasets. This process was repeated hundred of times and the final result is the mean of all the results obtained through individual trees. As the individual trees

are not pruned while construction so their variance is greatly improved so by averaging the results of individual their variance can be minimized by keeping the bias constant.

Iterative Impute [Buuren et al., (2011)]

It is a multiple imputation method, here imputing the missing values is by sequential processing of features. All the features are modeled as in terms of other features. The process of imputation starts with prediction of features values and these predicted values are useful for the next approximation of feature values.

Multi-Layer Preceptor [Murtagh et al., (1991)]

In general neural network work with two layers one input layer and another output layer, which can be useful for solving certain low dimensional problems. Multi-layer perceptron model is a three or more layers model with one input, another intermittent layer called hidden layer and third output layer. Each layer in multi-layer perceptron model associated with a non linear activation function. Out put of each layer is passed as input to the next immediate layer in the network. Multi-layer perceptron model are very useful for solving especially non-linear problems in the real world.

KNN Impute [Troyanskaya et al., (2001)]

K nearest neighbor is designed for classification and regression based on the type of the dataset. If the dataset is numeric in type then it leads to prediction problem. KNN imputation generally starts with finding the neighbor points which has minimum distance with given instance. K represents the number of neighbors required used for evaluation.

AdaBoost regression [Zhou (2012)]

It is and ensemble based boosting algorithm where the weak learning algorithm performance is improved for each iteration. This is a type of ensemble algorithm works on boosting the performance of weak learners. AdaBoost was originally proposed by Freund and Shapire for solving the binary classification problem. With this algorithm on each iteration of training instances the strongly predicted values get the lower weights and weakly predicted values get the larger weights. It mainly focuses on the earlier iteration and its corresponding performance value.

Mean, Median, Most-frequent and Constant Imputation [Little et al., (2014)]

Basic imputation techniques where each missing value in the column is calculated as mean of the column. In the same way missing values can be replaced by median of column and for most frequent imputation the missing values are replaced by the most frequent. Some

times constant values are imputed at missing values in the dataset which drives to constant imputation.

Lasso Regression [Hastie et al., (2017)]

Lasso regression is one similar to the ridge regression, here absolute value of coefficient is minimized rather than square of coefficient value. Both ridge and lasso regression are the form of regularization of linear regression. Bias and variance among the dataset can be minimized with help of introducing the feature complexities in the ordinary least square.

$$\min_{\beta} \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{p}|\beta_j| \qquad (3)$$

Above equation represents the Lagrange from of lasso regression which consisting of error (bias) term and complex term (variance). The x represents the data instance or training instances, β represents the coefficient of equation and λ represents penalty or complex term coefficient.

Contribution Work

Methodology

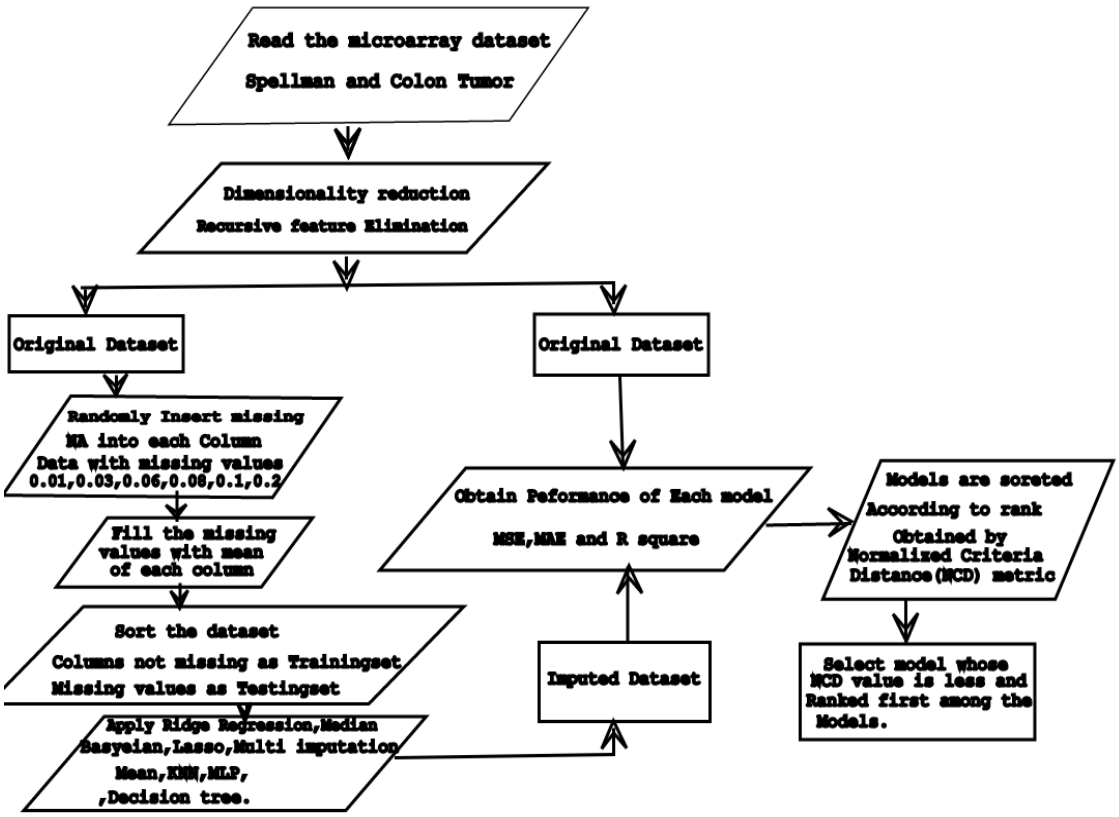

Fig. 1 Flow Chart of Methodology

1. Read the micro array gene expression dataset.
2. Perform the recursive feature elimination for dimensionality reduction.
3. Data obtained from step 2 is taken into two variables, these two same datasets are called original datasets.
4. Between these two original datasets, one is used for comparison and another is used for imputation procedure.
5. Randomly insert 0.01,0.03,0.06,0.8,0.1,0.2 percentage of missing values NA into each column of original dataset for imputation
6. Now insert the mean values of each column of missing values and sort the dataset such a way all complete values be the part of training instances and missing value columns will be treated as testing data.
7. Apply the Ridge regressor on the training set obtained in the step 6 and estimate the missing values through testing set.
8. Impute these estimated values into the positions of missing values in the dataset.
9. Now compare the performance of methodology by using the metrics MSE, MAE and R-square original dataset with imputed dataset from step 8.
10. Now assign the ranks to the models depending on performance values obtained through step 9 by using Normalized distance criteria (NCD).
11. The model with least NCD value is considered to be the best estimation model.

Ridge Regression: [Hastie et al., (2017)]

Regression is the way in which the target value is modeled to be a linear combination of the features, the model coefficients are calculated and must minimize the residual sum of squares.

The mathematical notation, if y is the predicted value then

$$y(w, x) = w_0 + w_1 x_1 + \cdots w_p x_p \qquad (4)$$

Fits a linear model with coefficients to minimize the residual sum of squares between the observed targets in the dataset and the targets predicted by the linear approximation. Mathematically it solves a problem of the form:

$$\min \|x_w - y\|^2 \qquad (5)$$

Ridge regression addresses some of the problems of ordinary least squares by imposing a penalty on the size of the coefficients. The ridge coefficients minimize a penalized residual sum of squares:

$$\min\|x_w - y\|^2 + \alpha\|w\|^2 \qquad (6)$$

The complexity parameter α controls the amount of shrinkage: the larger the value of α, the greater the amount of shrinkage and thus the coefficients become more robust to collinearity.

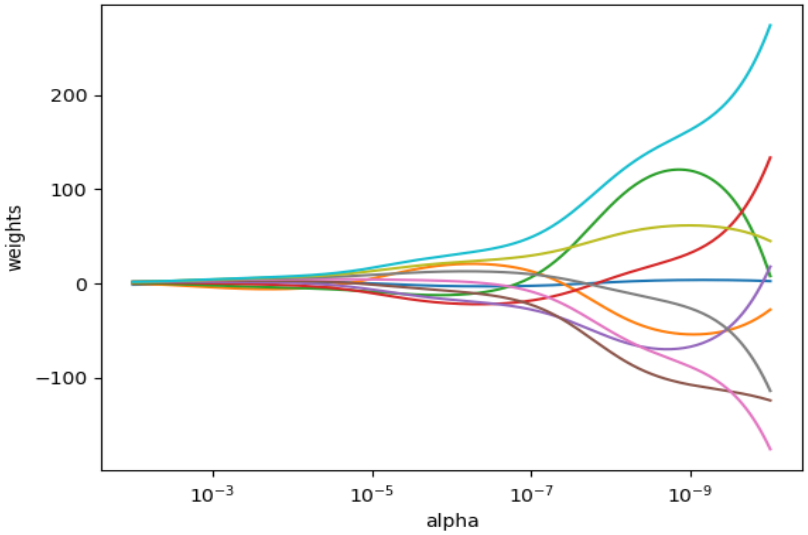

Fig. 2 Ridge Coefficients as a function of the regularization

Ridge regression is a powerful techniques used for creating models when data is with large number of features. It penalizes the magnitude of coefficients of features along with minimizing the error between predicted and actual observations, this technique is called regularization. Penalty is assigned in the following way.

Performs L2 regularization, i.e. adds penalty equivalent to square of the magnitude of coefficients

Minimization objective = LS Obj + α * (sum of square of coefficients)

Note that here LS Obj refers to least squares objective, i.e. the linear regression objective without regularization. Here, α (alpha) is a parameter which balances the amount of emphasis given to minimizing LS and minimizing sum of square of coefficients.

Performance Metrics

Performance of the proposed methodology and other methods of data estimation algorithms and their assessment is done with the metrics like MSE (mean square estimation), MAE (mean absolute error) and R square.

MSE (Mean Square Estimation)

Mean Squared error MSE is a good error metric for regression problems. It is also an essential loss characteristic for algorithm where fit is optimized using the least squares framing of a regression hassle. Least squares refer to minimizing the mean squared error between predictions and predicted values. The MSE is calculated because the mean or average of the squared differences among anticipated and anticipated target values in a dataset. The smaller values of MSE indicate the performance of the model is good.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i) \quad (7)$$

y_i: True value, $\hat{y}_i$: Predicted value and N represents total number of observations.

MAE (Mean Absolute Error)

Mean Absolute Error, or MAE, is a most used metric, just like RMSE (root mean square error), as its name suggests, the MAE score is calculated as the average of the absolute error values. Absolute or *abs()* is a mathematical function that makes a number positive. Therefore, the difference between an expected and predicted value may be positive or negative and is forced to be positive when calculating the MAE. MAE is less sensitive to outliers while compared with RMSE. If the value of MAE is low then it indicates that performance of the model is good.

The MAE can be calculated as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N} abs(y_i - \hat{y}_i) \quad (8)$$

R square (Coefficient of Determination)

R-squared (R^2) is a statistical measure that represents the variation in the target vector values. Generally, it gives the correlation among the observed values and predicted value. The value of R square varies between 0 and 1, if the R square value is 1 then model fits best for the training set or if R square value is 0 indicates the worst fit of model to training set.

$$R^2 = 1 - \frac{(y_i - \hat{y}_i)}{(y_i - \bar{\bar{y}}_i)} \quad (9)$$

$\bar{y}_i$: mean value of observed values.

Normalized Criteria Distance (NCD): [Pham (2013)]

With the availability of different prediction performance metrics, selecting the best model among different models is quite tough. Different performance measure has different impact

on the models. So, it is very not that easy to select the best model among existing prediction models. In this paper we use Normalized Criteria Distance (NCD) metric which assigns the ranks and select the model with best rank with assigned performance metrics. This metric was used by Hoang Pham to select the best software reliability growth model. Let m denotes the number of prediction models and c number of criteria are used for missing data estimation, CV_{ij} denotes the i^{th} criteria values for j^{th} prediction model. Where i = 1,2 3… m and j=1,2 3…c.

Now normalized Criteria distance value NCD_h, measures the distance of normalized value from the original for the h^{th} model and can be defined as

$$NCD_h = \sqrt{\left(\sum_{j=1}^{c}\left(\left(\frac{CV_{hj}}{\sum_{i=1}^{m} CV_{ij}}\right)^2\right)\right)} \qquad h = 1,2,3..m \qquad (10)$$

i={Extra-trees, Random-forest, iterative-impute, linear-SVR, Bayesian Ridge, KNN-impute, MLP-reg, Decision-tree-reg, AdaBoost, mean, lasso regression, median, most frequent,0 impute} and j={MSE,MAE,R-Square}.

Results and Discussion

A comparison of the proposed model with the previous models and studies is done using metrics MSE, MAE, R square and NCD.

Table III and IV shows the pattern of percentages of missing values in two datasets Spellman and Colon Tumor.

Table III Missing data pattern in Spellman dataset

Spellman															
	Columns														
%missing Data	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0.01	45	40	39	65	47	42	36	44	53	45	36	45	35	41	38
0.03	143	138	131	160	135	136	116	133	124	116	121	124	123	130	118
0.06	253	255	247	275	253	260	240	253	245	237	263	289	246	261	257
0.08	339	337	326	360	330	350	331	339	325	318	337	371	324	340	338
0.1	426	418	398	424	410	434	407	411	415	407	417	445	406	422	415
0.2	781	824	753	797	799	805	764	777	797	814	787	818	786	775	768

Table-IV Missing data pattern in Conol Tumor dataset

	Columns																			
% missing Data	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
0.01	24	17	18	35	22	13	15	21	22	22	18	22	15	18	16	18	19	21	20	23
0.03	58	56	55	74	67	61	60	64	52	54	54	54	56	54	49	58	63	61	62	70
0.06	101	111	95	126	124	128	117	122	105	108	113	127	118	112	108	111	126	128	125	124
0.08	136	136	133	171	158	172	166	161	145	141	150	168	151	144	143	153	153	161	159	151
0.1	185	185	165	207	204	210	202	213	181	183	186	201	174	177	168	198	196	196	190	182
0.2	368	373	340	400	360	388	354	371	366	368	350	367	361	348	346	368	357	386	345	350

Table V and VI shows the various models and their corresponding performance metric values for two datasets Spellman, and Conol Tumor. The last columns in the Table Vand Table VI shows the NDC (Normalized criteria distance) for various models. From the Table V the NCD value for Ridge regressor is 0.36254which is lower than any other model. From the Table VI the NCD value for Conol Tumor dataset for Ridge regressor is 0.38738which is a lower value compared with other models. The model with lowest NCD value is higher rank among the other models. From the Table V and VI NCD value for the Ridge regressor with proposed methodology got highest rank among other models.

Table V Performance comparison for various models for spellman dataset

Spellman				
Name	MSE	MAE	R Square	Dk
Ridge regression	0.000206	0.0026477	0.967382	0.36254
Random-forest	0.00021067	0.0026512	0.966453	0.36271
iterative impute	0.000222	0.002785	0.965128	0.36537
MLP reg	0.00031617	0.00333	0.94982	0.37705
KNN impute	0.00032883	0.003506	0.948419	0.38035
AdaBoost	0.00032517	0.0036343	0.946481	0.38201
Decision-tree reg	0.00042033	0.0038203	0.937774	0.38838
Median	0.0004995	0.0044248	0.916713	0.39896
Mean	0.000498	0.0044352	0.916932	0.39907
lasso regression	0.000498	0.0044352	0.916932	0.39907
most frequent	0.00053333	0.0045715	0.912694	0.40226
0 impute	0.02359967	0.0408977	0.244368	1.1749

Table VI Performance comparison for various models for Conol Tumor

Conol Tumor				
Model	MSE	MAE	R Square	Dk
Ridge regression	8.1167E-05	0.001194	0.988038	0.38738
iterative impute	9.1333E-05	0.0012833	0.986749	0.39348
KNN impute	0.00011217	0.0012558	0.983438	0.39729
MLP reg	0.00011783	0.0014395	0.982523	0.40587
Decision-tree reg	0.00015133	0.0015327	0.978345	0.41719
AdaBoost	0.00023333	0.003452	0.965681	0.50032
Median	0.00060817	0.0029098	0.902512	0.55017
Mean	0.00055733	0.0034742	0.909441	0.55768
lasso regression	0.00055733	0.0034742	0.909441	0.55768
0 impute	0.000779	0.0039888	0.880789	0.60773
most frequent	0.00166233	0.0074938	0.830536	0.80879

Conclusion

Missing data is an important issue especially in case of data analysis. Missing data or values can degrade the performance of learning algorithms, therefore it is highly important to estimate and imputer the missing values before processing. Micro array gene expression data is biological data; drive the characteristics of genes and their corresponding values under experimental condition. Now these expression data have huge dimensionality, can degrade the performance of machine learning or any algorithm. In this paper a new methodology with Recursive Feature Elimination for dimensionality reduction combined with Ridge Regression a regularization based algorithm for missing values estimations is proposed. Metrics like MSE, MAE and NCD are used to evaluate the performance. A comparative analysis is done with other imputation techniques and is observed an improved performance with proposed methodology.

References

Acuña, E., & Rodriguez, C. (2004). The treatment of missing values and its effect on classifier accuracy. *Classification, Clustering, and Data Mining Applications*, 639-647. https://doi.org/10.1007/978-3-642-17103-1_60

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., & Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, *96*(12), 6745-6750. https://doi.org/10.1073/pnas.96.12.6745

Armstrong, S.A., Staunton, J.E., Silverman, L.B., Pieters, R., Den Boer, M.L., Minden, M.D., Sallan, S.E., Lander, E.S., Golub, T.R., & Korsmeyer, S.J. (2001). MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genetics*, *30*(1), 41-47. https://doi.org/10.1038/ng765

Awad, M., & Khanna, R. (2015). *Efficient learning machines: Theories, concepts, and applications for engineers and system designers*. Apress.

Bishop, C.M. (2006). *Pattern recognition and machine learning*. Springer Verlag.

Breiman, L. (2001). *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/a:1010933404324

Breiman, L., Friedman, J., Stone, C.J., & Olshen, R. (1984). *Classification and regression trees*. Chapman and Hall/CRC.

Brown, M.L., & Kros, J.F. (2003). Data mining and the impact of missing data. *Industrial Management & Data Systems*, *103*(8), 611-621. https://doi.org/10.1108/02635570310497657

Buuren, S. V., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3). https://doi.org/10.18637/jss.v045.i03

Celton, M., Malpertuy, A., Lelandais, G., & De Brevern, A. G. (2010). Comparative analysis of missing value imputation methods to improve clustering and interpretation of microarray experiments. *BMC Genomics*, *11*(1). https://doi.org/10.1186/1471-2164-11-15

Eekhout, I., De Boer, R.M., Twisk, J.W., De Vet, H.C., & Heymans, M.W. (2012). Missing data. *Epidemiology*, *23*(5), 729-732. https://doi.org/10.1097/ede.0b013e3182576cdb

Farhangfar, A., Kurgan, L., & Dy, J. (2008). Impact of imputation of missing values on classification error for discrete data. *Pattern Recognition*, *41*(12), 3692-3705. https://doi.org/10.1016/j.patcog.2008.05.019

Hastie, T., Tibshirani, R., & Friedman, J. (2013). *The elements of statistical learning: Data mining, inference, and prediction*. Springer Science & Business Media.

Hoheisel, J.D. (2006). Microarray technology: Beyond transcript profiling and genotype analysis. *Nature Reviews Genetics*, *7*(3), 200-210. https://doi.org/10.1038/nrg1809

Kuhn, M., & Johnson, K. (2019). *Feature engineering and selection: A practical approach for predictive models*. CRC Press.

Li, H., Zhao, C., Shao, F., Li, G., & Wang, X. (2015). A hybrid imputation approach for microarray missing value estimation. *BMC Genomics*, *16*(S9). https://doi.org/10.1186/1471-2164-16-s9-s1

Little, R.J., & Rubin, D.B. (2014). *Statistical analysis with missing data*. John Wiley & Sons.

Multiple imputation for Nonresponse in surveys. (1987). *Wiley Series in Probability and Statistics*. https://doi.org/10.1002/9780470316696

Murtagh, F. (1991). Multilayer perceptrons for classification and regression. *Neurocomputing*, *2*(5-6), 183-197. https://doi.org/10.1016/0925-2312(91)90023-5

Pham, H. (2013). A new software reliability model with vtub-shaped fault-detection rate and the uncertainty of operating environments. *Optimization*, *63*(10), 1481-1490. https://doi.org/10.1080/02331934.2013.854787

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581-592. https://doi.org/10.1093/biomet/63.3.581

Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., & Futcher, B. (1998). Comprehensive identification of cell cycle–regulated genes of the Yeast Saccharomyces cerevisiae by Microarray hybridization. *Molecular Biology of the Cell*, *9*(12), 3273-3297. https://doi.org/10.1091/mbc.9.12.3273

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R.B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, *17*(6), 520-525. https://doi.org/10.1093/bioinformatics/17.6.520

Christobel, A., & Siva Prakasam, P. (2012). The negative impact of missing value imputation in classification of diabetes dataset and solution for improvement. *Journal of Computer Engineering*, *7*(4), 16-23. https://doi.org/10.9790/0661-0741623

Yoon, D., Lee, E., & Park, T. (2007). Robust imputation method for missing values in microarray data. *BMC Bioinformatics*, *8*(S2). https://doi.org/10.1186/1471-2105-8-s2-s6

Zhou, Z. (2012). *Ensemble methods: Foundations and algorithms*. CRC Press.

Author's Profile

K. Ashfaq Ahmed. received Master of Technology in Computer Science from Jawaharlal Nehru Technological University, Anantapuram, A.P, India. With more than 15yrs of service as a faculty of Computer Science at colleges in India and abroad. Areas of research are Artificial Intelligence, Machine Learning and Data Analytics, Published several research papers in various national, international conferences and journals.

Dr. Shaheda Akthar received Bachelor of Computer Science and Master of Computer Science from Acharya Nagarjuna University, M.S from B.I.T.S Pilani. Ph.D from Acharya Nagarjuna University. Presently working as Faculty of Computer Science, Government College for Women (A), Guntur, A.P, India and Research Director for Department of Computer Science & Engineering, Acharya Nagarjuna University, Guntur, A.P, India. Areas of interest are Software Engineering, reliability and quality control, Software Architecture Recovery. Machine Learning and Data Mining. Published more than 35 research papers in various international journals.